РУБРИКА

«ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»

РАЗРАБОТКА СИСТЕМЫ АНАЛИЗА ТОНАЛЬНОСТИ СООБЩЕНИЙ В СОЦИАЛЬНОЙ СЕТИ TWITTER

Авдеева Татьяна Геннадьевна
студент, кафедра ИСОиП (филиал) ДГТУ,
РФ, г. Шахты
E-mail: tan.awdeewa2011@yandex.ru

THE SYSTEM DEVELOPMENT FOR TWITTER SENTIMENT ANALYSIS

Tatyana Avdeeva
student, institute of service and business (branch) DSTU
Russia, Shakhty

АННОТАЦИЯ

В статье описывается разработка программы для анализа тональности текста в социальной сети Twitter на основе методов машинного обучения. Проведен анализ существующих методов определения тональности текста и векторных моделей представления текста. Было реализовано три алгоритма классификации текстов по тональности – наивный Байесовский алгоритм, случайные леса и логистическая регрессия. В ходе тестирования алгоритмов при помощи ROC – кривой, был определен наиболее эффективный.

ABSTRACT

The article describes the development of a program for twitter sentiment analysis based on machine learning methods. The analysis of current methods for determining sentiment analysis and vector space models of text presentation were provided. Three algorithms for the classification of texts by tonality were implemented: naive Bayesian algorithm, random forests, and logistic regression. During the testing of algorithms using the ROC curve, the most effective one was determined.

Ключевые слова: машинное обучение, векторная модель, наивный Байесовский алгоритм, случайные леса, логистическая регрессия.

Keywords: machine learning, vector space model, naive Bayesian classifier, random forests, logistic regression.

Введение. В настоящее время активно развивается такие сферы жизни человека как обмен информации средствами сети Интернет. Увеличение количества социальных сетей, блогов, форумов, веб – ресурсов приводит к появлению таких задач, как анализ текстов пользователей по различным вопросам (отношения к событиям, отзывы о товарах, услугах, мнения о высказываниях, оценка пользователей по отношению к другим людям). Одной из основных проблем при анализе текста является анализ тональности текста – это быстро развивающееся направление компьютерной лингвистики, основной задачей которого является выявление в документе эмоционально окрашенной лексики и эмоциональной оценки объектов автором.

Процесс разработки системы автоматического анализа тональности сообщений в рамках настоящего исследования, схож с процессом разработки любой другой программы на основе методов машинного обучения.

Анализировать эмоциональную окраску твиттов особенно сложно из-за ограничения на размер твитта – всего 140 символов. Отсюда и специальный синтаксис, и нестандартные аббревиатуры, и неправильно построенные предложения. Классический подход определения тональности сообщений здесь не подходит.

Методология анализа. Используются различные наборы функций и машинного обучения классификаторов для определения лучшей комбинации для анализа настроений в Twitter. Также проводилась предварительная обработка, такая как – знаки препинания, смайлики, с отдельными твитами. Исследовались следующие характеристики – unigrams, биграмм, триграмм и отрицание обнаружения. Классификатор был подготовлен с помощью различных алгоритмов машинного обучения – наивного Байесовского классификатора, деревья решений и логистической регрессии. На рисунке 1 представлено схематическое представление методологии.

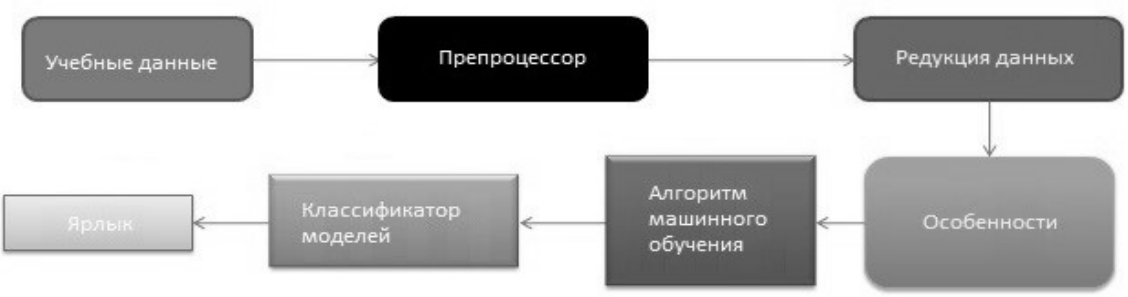

Рисунок 1. Схематическое представление методологии

«Twitter – социальная сеть для публичного обмена сообщениями при помощи веб – интерфейса, SMS, средств мгновенного обмена сообщениями или сторонних программ – клиентов». [4] Для использования твитов во внешних программных продуктах использовался Twitter API.

С помощью Twitter API можно:

- получать свои и чужие ленты твитов;
- добавлять новые твиты;
- удалять и обновлять старые записи;
- получать только подходящие по параметрам твиты;
- получать списки наших читателей и их друзей;
- изменять настройки нашей учетной записи;
- блокировать спам.

В качестве языка разработки был выбран Python. «Python – это высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. В то же время стандартная библиотека включает большой объём полезных функций». [2, с. 302]

В качестве среды разработки была выбрана Jupyter Notebook. Jupyter Notebook – это веб-среда, которая позволяет выполнять интерактивные вычисления.

Предлагаемое решение. Для определения наиболее эффективного алгоритма машинного обучения для автоматического анализа тональности текстовых сообщений в Twitter были разработаны два алгоритма. Первый – алгоритм бинарной классификации текстовых сообщений, который взят в данном проекте в качестве эталона для сравнения. На рисунке 2 представлена схема алгоритма. К каждому этапу приведены обоснования его использования. Данный алгоритм предполагает разделение выборки сообщений только на два класса: «позитивные» и «негативные».

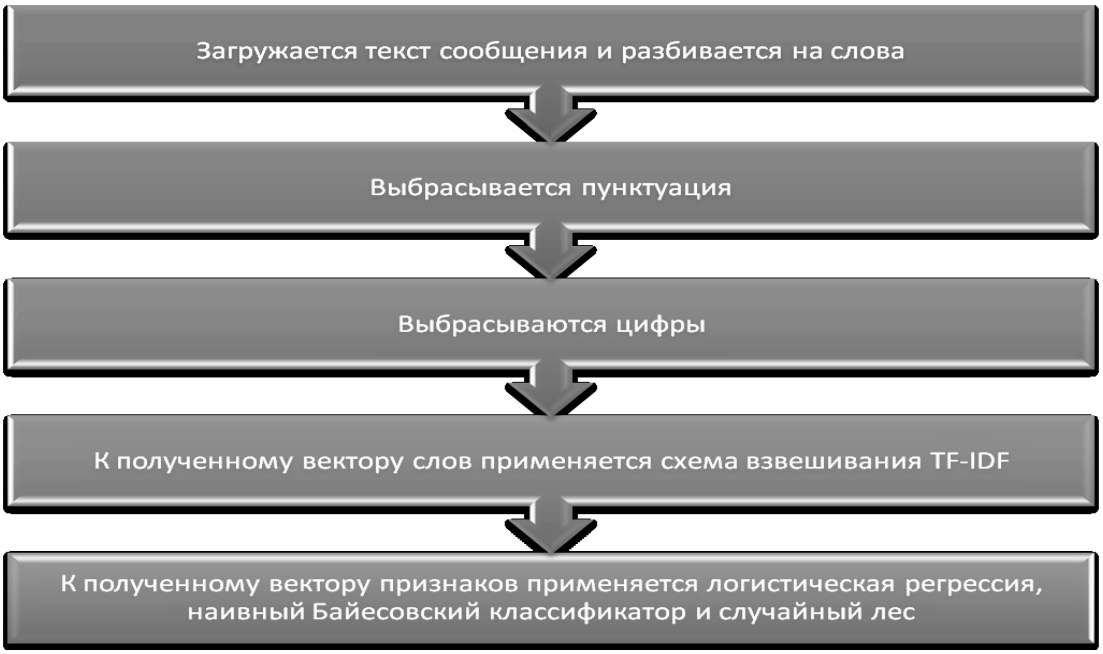

Рисунок 2. Общий вид алгоритма Бинарной классификации

1. Загружаем текст сообщения и разбиваем на слова, которые потом приводим к нижнему регистру.
2. Выбрасывается пунктуация.
3. Выбрасываются цифры.

4. К полученному вектору слов применяется схема взвешивания TF-IDF.

5. К полученному вектору признаков применяется логистическая регрессия, наивный Байесовский классификатор и случайный лес.

Для предобработки и получения признаков использовалась библиотека машинного обучения scikit-learn.

Для тестирования точности алгоритмов классификации, был использован метод перекрестной проверки (скользящий контроль, cross-validation). «В рамках этой процедуры фиксируется некоторое множество разбиений исходной выборки на две подгруппы: обучающую и контрольную». [3, с. 5] Для каждого разбиения выполняется настройка алгоритма по обучающей подгруппе, затем оценивается его эффективность на векторах контрольной подгруппы. Оценкой перекрестной проверки называется средние по всем разбиениям величины точности и полноты на контрольных подгруппах. Если объекты выборки независимы, то средние значения оценок эффективности перекрестной проверки дадут несмещенные оценки эффективности.

На рисунке 3 показана точность классификации трех алгоритмов на тестовых данных.

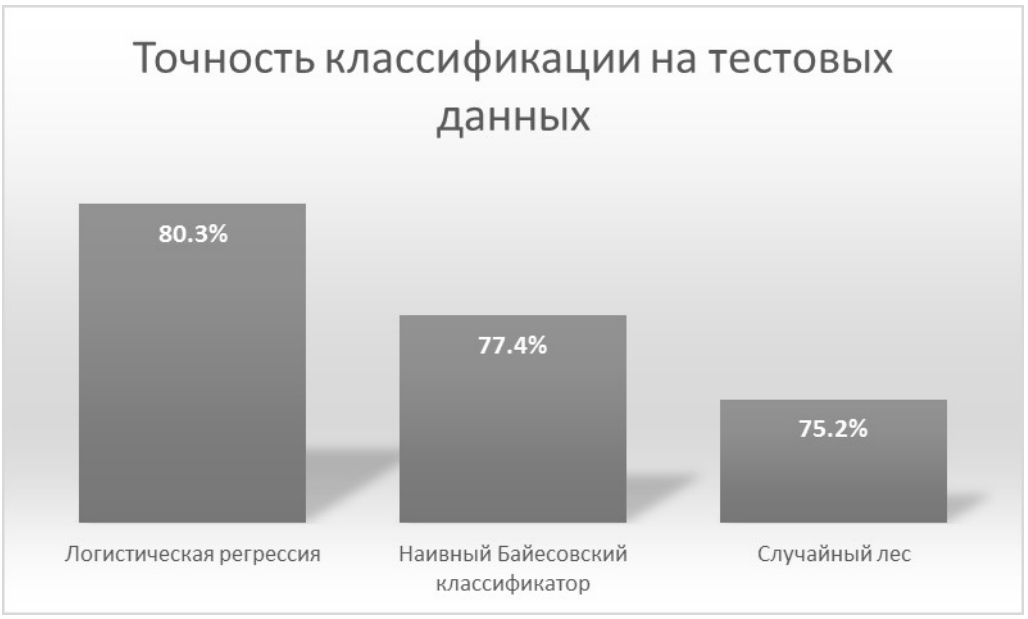

Рисунок 3. Точность классификации

Как видно из рисунка 3, что логистическая регрессия имеет наибольшую точность.

Эмоциональная окраска твиттов может быть не только положительной или отрицательной, но также всегда присутствуют твитты которые вообще никак не окрашены эмоционально – они нейтральны или нерелевантные и просто содержат информацию. Для классификации текстовых сообщений на три класса был применен иерархический подход. Подход заключается в отделении на первой стадии тональных сообщений от нейтральных, а на второй – в разделении тональных на позитивные и негативные, как представлено на рисунке 4.

Рисунок 4. Общий вид алгоритма Иерархической классификации

В представленном алгоритме иерархической классификации поиск нейтральных твиттов осуществляется на основе корпуса нейтральных и эмоционально окрашенных слов Liu and Hu opinion lexicon.

На рисунке 5 показана точность классификации трех типов классификаторов на тестовых данных.

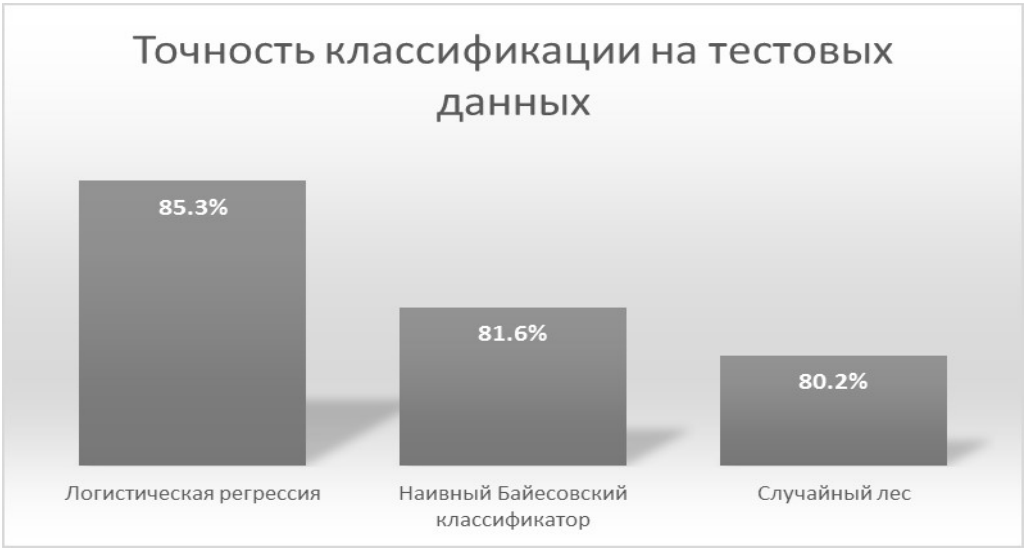

Рисунок 5. Точность классификации

Как видно из рисунка 5, что логистическая регрессия имеет наибольшую точность классификации.

Сравнивая два типа классификации (рисунок 6) можно сделать вывод, что иерархическая классификация в среднем на 5% лучше бинарной.

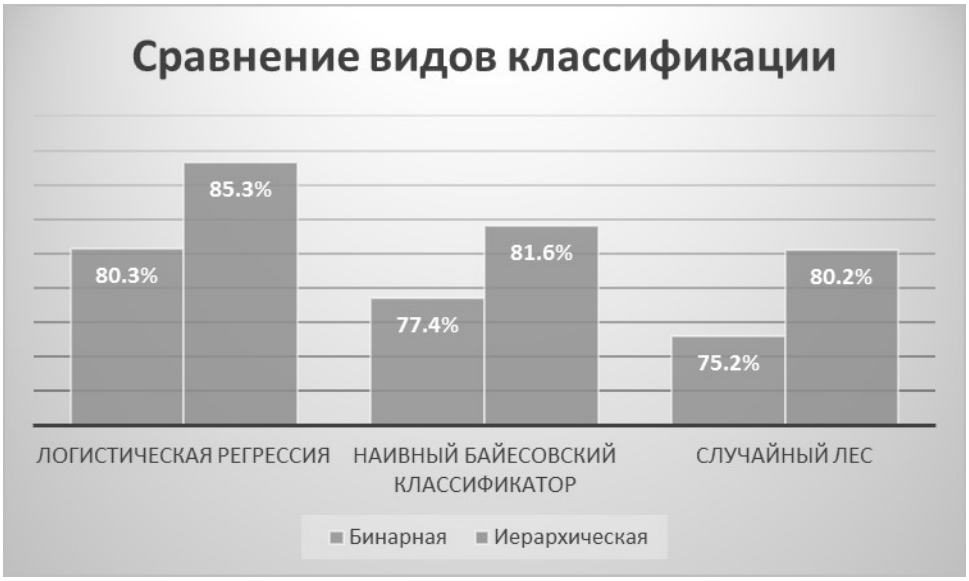

Рисунок 6. Сравнение видов классификации

На сегодняшний день существует очень мало публичных русскоязычных коллекций сообщений, которые можно было бы использовать для решения задачи классификации отзывов на три класса (положительные, отрицательные, нейтральные), и не обнаружено ни одной русскоязычной публичной коллекции постов микроблогов.

Для анализа был использован корпус из 230 000 автоматически размеченных (с точностью более 82%) твитов, который есть в открытом доступе. В качестве алгоритма для подбора типа классификатора был выбран алгоритм бинарной классификации текстовых сообщений.

На рисунке 7 показана точность классификации трех типов классификаторов на тестовых данных для русскоязычного сегмента твиттера.

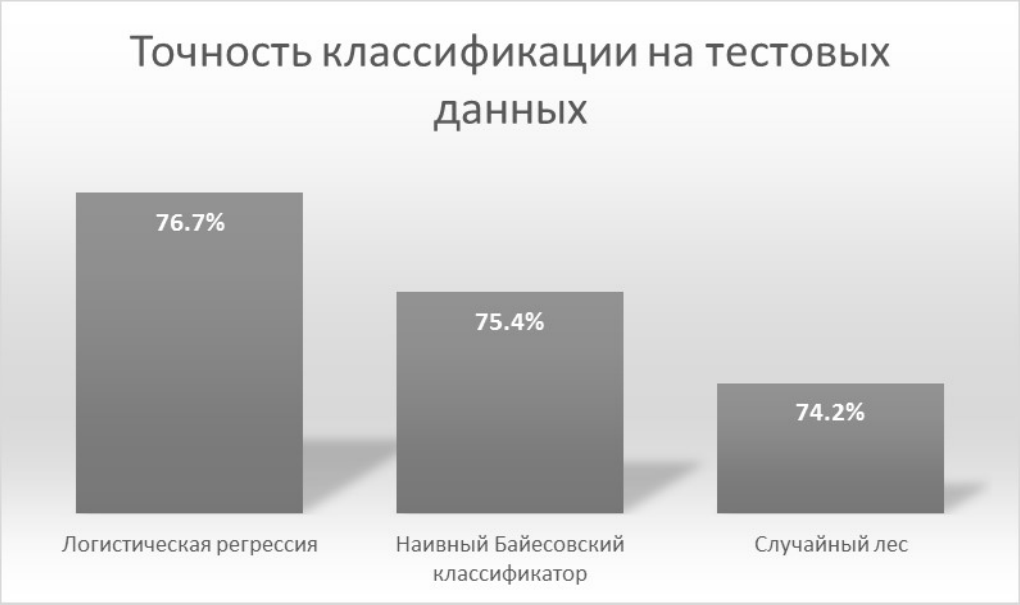

Рисунок 7. Точность классификации

Как видно из результатов исследования, точность классификации для русскоязычного сегмента твиттера уменьшилась, так как русский язык сложен в морфологическом и лексическом плане.

Выводы. В данной статье была рассмотрена реализация наиболее популярных алгоритмов классификации, а также определён наиболее эффективный алгоритм машинного обучения для анализа тональности текстовых сообщений в Twitter. Описывается среда разработки и язык программирования. Так же был представлен алгоритм для анализа англоязычных текстовых сообщений и русскоязычных текстовых сообщений в Twitter.

Список литературы:

1. Храмов Д. Использование Twitter API для сбора данных. URL: http://dkhramov.dp.ua/images/edu/Stu.WebMining/ch17_twitter.pdf (дата обращения 30.03.2017).
2. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. // «Машинное обучение», Изд. 2е – 2016 -302 с.
3. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. — Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — 5 с.